

PR #2252 完整报告

THUDM/slime

[release] bump to v0.3.1

合并时间: 2026-08-06 19:08

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2252>

执行摘要

- 一句话: 发布 v0.3.1, 集成 SGLang/Megatron 补丁与构建修复
- 推荐动作: 值得精读, 尤其适合维护 docker 补丁和发布流程的团队。关注补丁的组织方式、版本兼容策略, 以及 reloadable process group 销毁顺序的修复思路。

功能与动机

作为 release PR, 目标是发布 v0.3.1。提交历史显示发布前修复了多个阻断项, 如 'Fix CUDA headers in conda builds'、'Fix conda build and Qwen3 test memory'、'Drop pre-0.3 sglang-router compatibility', 确保 Docker 镜像与核心链路可用。

实现拆解

1. 版本号更新: 将版本号改为 0.3.1 (commit "change version to 0.3.1")。
2. 引入 SGLang 与 Megatron 补丁: 新增 docker/patch/v0.5.15.post1 下的 5 个补丁, 覆盖 PD 分解超时与内存释放 (sglang.patch)、top_p 采样支持 (sglang-top_p.patch)、release_hicache 缓存释放 (sglang-release_hicache.patch)、权重拉取端点 (sglang-pull_weights.patch) 以及 Megatron checkpoint 加载容错 (megatron.patch)。这些补丁与 SGLang v0.5.15.post1 配套, 是发布的基础。
3. 移除旧版 sglang-router 兼容: 在 slime/backends/sglang_utils/sglang_engine.py 中删除了对 router $\leq 0.2.1$ 的 add_worker 路径和低版本 remove_worker 逻辑, 统一使用新 API (/workers 端点), 简化了注册与关闭流程。
4. 修复外部 rollout 并行配置: 在 slime/backends/sglang_utils/external.py 中为 ExternalEngineInfo 增加 parallel_config 字段并传递给 rollout, 同时新增测试 tests/test_external_slang_engines.py 验证。
5. 修复 reloadable process group 销毁顺序: 在 slime/utils/reloadable_process_group.py 中新增 invalidate_process_groups 静态方法, 并在 _destroy_default_nccl_process_group 中先销毁 WORLD 再清空所有子组包装器句柄, 解决 PP=4 拓扑下 rank 间销毁顺序不一致导致的死锁。
6. 构建与 CI 修复: 修复 conda 构建中的 CUDA 头文件问题、pin torchvision 版本、降低 Qwen3 测试内存占用, 并修剪 dense parallel CI 矩阵以加快验证。

关键文件:

- docker/patch/v0.5.15.post1/sglang.patch (模块 引擎补丁; 类别 infra; 类型 infrastructure; 符号 DecodeTransferQueue, release_memory_occupation, resume_memory_occupation, get_weights_by_name) : 新增 SGLang 核心补丁, 覆盖 PD 分解超时、内存释放、权重版本等, 与 SGLang v0.5.15.post1 配套, 是本次发布的基础。
- docker/patch/v0.5.15.post1/megatron.patch (模块 训练补丁; 类别 infra; 类型 infrastructure; 符号 _validate_global_shapes, TEGroupedLinear, _encode_extra_state, _postprocess) : 新增 Megatron 补丁, 修改 checkpoint 加载容错、DDP 参数等, 解决发布前遇到的训练侧兼容问题。
- docker/patch/v0.5.15.post1/sglang-top_p.patch (模块 采样补丁; 类别 infra; 类型 infrastructure; 符号 get_dsa_seed_metadata_dim, _forward_ascend_backend, _attach_logprobs_to_output, get_token_ids_logprobs_raw) : 新增 top_p 采样补丁, 为 PD 分解场景传递 top_p token ids, 扩展输出能力。
- docker/patch/v0.5.15.post1/sglang-release_hicache.patch (模块 缓存补丁; 类别 infra; 类型 infrastructure; 符号 write_backup, get_height, init_kv_buffer, get_ksize_per_token) : 新增 release_hicache 补丁, 支持在权重卸载时释放 KV cache 内存, 配合 --release-train 使用。
- docker/patch/v0.5.15.post1/sglang-pull_weights.patch (模块 权重补丁; 类别 infra; 类型 infrastructure; 符号 update_weight_version, SlowDownReqInput, check_weights) : 新增 pull_weights 补丁, 为 SGLang 增加权重拉取端点, 配合磁盘级 delta 同步。
- slime/backends/sglang_utils/sglang_engine.py (模块 SGLang 引擎; 类别 source; 类型 dependency-wiring; 符号 _register_to_router, shutdown) : 移除旧版 sglang-router 兼容, 统一注册与关闭流程, 影响所有 SGLang 引擎的启停。
- slime/utils/reloadable_process_group.py (模块 进程组重载; 类别 source; 类型 core-logic; 符号 invalidate_process_groups) : 修复销毁 / 重载 WORLD 时子组句柄失效顺序, 新增 invalidate_process_groups, 解决 PP=4 死锁。
- slime/backends/sglang_utils/external.py (模块 外部引擎; 类别 source; 类型 core-logic; 符号 parallel_config) : 为外部 rollout 引擎增加 parallel_config 信息, 确保并行配置正确传递。

关键符号: invalidate_process_groups, _destroy_default_nccl_process_group, _register_to_router, start_external_rollout_servers, release_memory_occupation

关键源码片段

slime/utils/reloadable_process_group.py

修复销毁 / 重载 WORLD 时子组句柄失效顺序, 新增 invalidate_process_groups, 解决 PP=4 死锁。

```
@staticmethod
def invalidate_process_groups():
    """Drop handles after destroying WORLD, which already shut down every subgroup."""
    pid = os.getpid()
    for reloadable_group in ReloadableProcessGroup.GROUPS.get(pid, []):
```

```

reloadable_group.group = None

# 在销毁 NCCL WORLD 时，先销毁 WORLD 让 PyTorch 按全局顺序关闭所有
# 已注册的 NCCL/Gloo 后端，再调用 invalidate_process_groups 丢弃包装器
# 中的陈旧句柄，避免 PP=4 场景下各 rank 子组销毁顺序不一致造成死锁。
def _destroy_default_nccl_process_group() -> None:
    state = default_process_group_states.get(os.getpid())
    if state is None or state.nccl_world_destroyed or not _uses_nccl(state.backend):
        return

    dist.barrier(group=get_gloo_group())
    dist.destroy_process_group()
    ReloadableProcessGroup.invalidate_process_groups()
    set_gloo_group(None)

    _new_default_process_group(state, backend="gloo")
    set_gloo_group(_get_default_group())
    state.nccl_world_destroyed = True
    logger.info(
        "Destroyed default %s WORLD process group and initialized a temporary Gloo WORLD
        (generation %s)",
        state.backend,
        state.generation,
    )

```

评论区精华

该 PR 没有 review 评论，但 13 个提交迭代反映了发布前的主要修复点，重点集中在构建、兼容性和进程组销毁顺序。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 补丁漂移风险：大量 SGLang/Megatron 补丁绑定 v0.5.15.post1，上游更新后补丁可能冲突，维护成本高。
2. 旧环境破坏性变更：sglang_engine.py 移除旧版 router 兼容，若部署环境仍使用 slang-router ≤0.2.1，将无法注册 worker 或关闭时失败。
3. 进程组顺序变更影响：reloadable process group 的销毁逻辑改变，依赖旧行为的代码可能在多 rank 场景出现新死锁。
4. 构建修复未完全验证：conda 构建修复、torchvision pin 等仅在本 PR 中提交，未充分覆盖所有 GPU 驱动组合，可能影响镜像可移植性。- 影响：影响 Docker 镜像构建、SGLang 引擎注册与关闭、外部引擎发现、Megatron checkpoint 加载、分布式进程组重载，以及 CI 测试矩阵。对使用者而言，需升级至 slang-router ≥0.3.0，并重新构建镜像以获取修复。- 风险标记：大量外部补丁引入，依赖版本绑定，移除旧版 router 兼容，破坏旧环境，进程组销毁顺序变更，存在新死锁风险，构建修复较多，需验证

关联脉络

- PR #2228 [docker] upgrade slang to v0.5.15.post1: 本次发布将 SGLang v0.5.15.post1 补丁版本化, 依赖该升级。
- PR #2181 [3/n] Disaggregated rollout: engine-side /pull_weights: pull_weights 补丁被本次发布纳入, 形成完整权重同步链路。
- PR #2208 Support reloading the default process group: reloadable process group 的销毁顺序修复与本次发布集成。
- PR #2180 Add --release-train: release_hicache 补丁与 --release-train 功能相关。